Count Regression

Andy Grogan-Kaylor

7 Jul 2020

# Key Concepts and Commands

* formulas are our friends
* Think about OR’s, pred. prob., non-linearity
* So much of categorical data analysis depends upon arguments for “functional form”. When do we think these arguments are valid?

# The Count



The Count and Friends

# Historical Examples of Count Data <U+0001F434> <U+260E><U+FE0F> <U+0001F3E5>

* <U+0001F434> Deaths from horsekicks in the Prussian Army
* <U+260E><U+FE0F> Calls to a call center (business, crisis hotline, etc.)
* <U+0001F3E5> Arrivals at the Emergency Room

# Other Canonical Examples of Count Data <U+0001F38B> <U+0001F48A>

* <U+0001F331> <U+0001F333> <U+0001F331> Plants / trees in a field
* <U+0001F927> <U+0001F48A> <U+0001F637> Cases of disease / unit area

# Poisson Distribution

## Theorizing about the Poisson <U+0001F41F> <U+0001F41F> <U+0001F41F> <U+0001F334> <U+0001F332> <U+0001F333>

The Poisson distribution is a modified form of the binomial distribution that is useful for the analysis of phenomena characterized by discrete, rare events. For example, in a study of the distribution of a rare plant among a number of quadrats, a majority of plots may contain no specimens, a smaller number a single plant, and still smaller numbers two, three, or more plants. If a single plant per quadrat is expected, the mean µ = 1 and the “0” and “1” classes occur at 37% each, the “2” class at 18%, the “3” class at 6%, and larger classes take up the remaining 2%. The characteristic test for a Poisson is that the variance equals the mean, which in the plant example means that the rare plant is randomly distributed. Note, that, In the distributions above, as the mean µ increases towards 10, the distribution approaches normality.

The Poisson may be used to evaluate whether events occur independently in time as well as space. In a classic example, Bortkiewicz (1898) studied the distribution of 122 men kicked to death by horses among ten Prussian army corps over 20 years. In most years in most corps, no one dies from being kicked; in one corp in one year, four men were kicked to death. Does this mean something was amiss in this particular corp? Analysis indicates that the observed frequencies conform quite closely to the expected Poisson frequencies, and that the mean and variance are almost identical, as expected. The corp was just “unlucky” in the sense that it is in the extreme tail of an ordinary run of events.

<http://www.mun.ca/biology/scarr/smcPoisson_distributions.html>

## Reprise of Normal Distribution

Normal distribution:

Standardized Normal Distribution:

2 parameters:

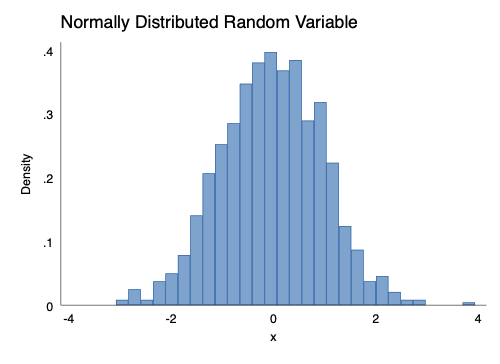
. clear all // clear all for simulated data

. set obs 1000 // number of observations  
number of observations (\_N) was 0, now 1,000

. generate x = rnormal() // normally distributed random variable

. histogram x, title("Normally Distributed Random Variable") scheme(burd)  
(bin=29, start=-3.0796595, width=.24185289)

. graph export myhistogram.png, width(500) replace  
(file myhistogram.png written in PNG format)



histogram of random normal variable

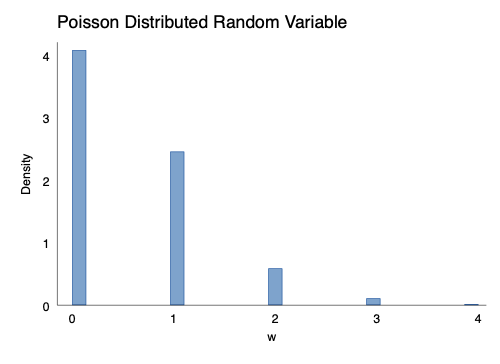
## Poisson Distribution

A Poisson with large lambda looks very similar to a normal distribution.

. generate w = rpoisson(.5)

. histogram w, title("Poisson Distributed Random Variable") scheme(burd)  
(bin=29, start=0, width=.13793103)

. graph export myhistogram2.png, width(500) replace  
(file myhistogram2.png written in PNG format)



histogram of random Poisson variable

## Poisson is the Natural Form for Observations Distributed in Time or Space <U+0001F3ED> <U+0001F3E2> <U+0001F3E8> <U+23F3> <U+23F3> <U+23F3>

is both mean and variance.

. clear all

. set obs 20  
number of observations (\_N) was 0, now 20

. generate field = \_n // field number

. generate mycount = rpoisson(1)

. expand mycount // create individual observations  
(6 zero counts ignored; observations not deleted)  
(8 observations created)

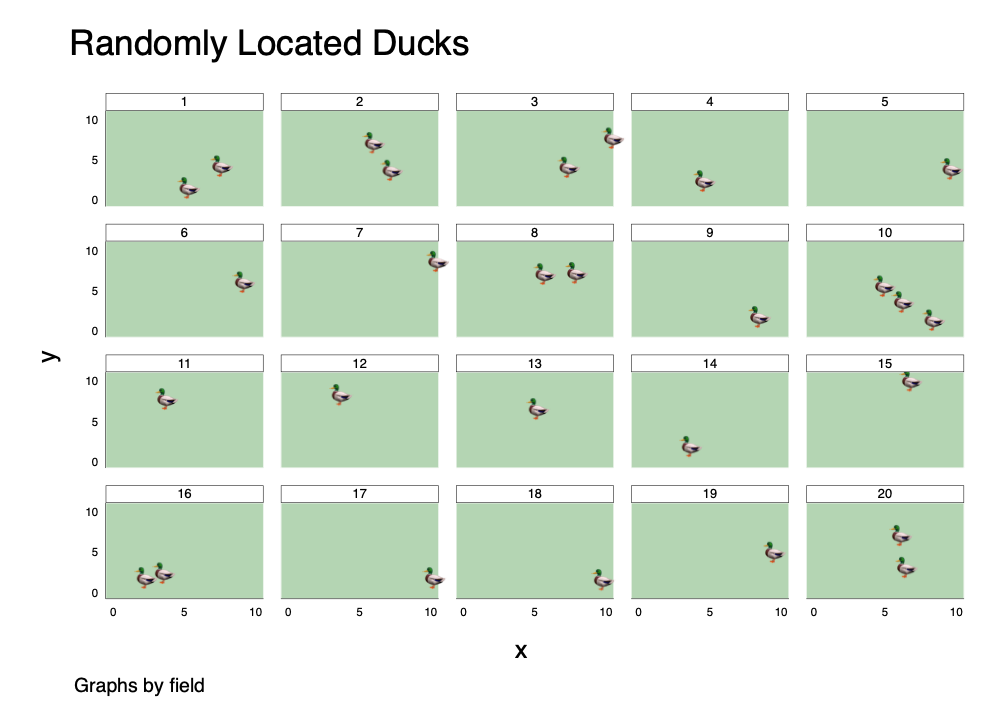
. generate x = runiform(1,10) // random x coordinate

. generate y =runiform(1,10) // random y coordinate

. generate mylabel = "🦆"

. twoway scatter y x, ///  
> by(field, title("Randomly Located Ducks")) ///  
> mlab(mylabel) mlabsize(vlarge) ///  
> msymbol(none) ///  
> legend(order(1 "🦆 Duck")) ///  
> scheme(burd) plotr(fcolor(olive\_teal))

. graph export ducks.png, width(1000) replace  
(file ducks.png written in PNG format)



Randomly Located Ducks

. generate t = runiform(1,10) // random time coordinate

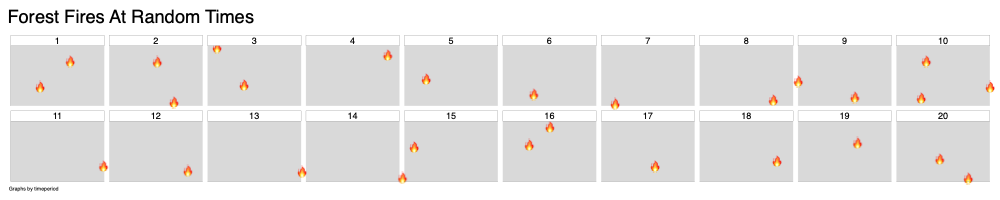
. generate y2 = runiform(1,2) // random y coordinate

. generate mylabel2 = "🔥"

. generate timeperiod = field

. twoway scatter y2 t, ///  
> by(timeperiod, title("Forest Fires At Random Times", size(vhuge)) cols(10)) ///  
> ytitle("", size(zero)) ylabel(none) xtitle("", size(zero)) xlabel(none) ///  
> subtitle(, size(vhuge)) ///  
> mlab(mylabel2) mlabsize(vhuge) ///  
> msymbol(none) ///  
> legend(order(1 "🔥 Forest Fire")) ///  
> scheme(burd) plotr(fcolor(gs14)) ///  
> xsize(5) ysize(1)

. graph export fires.png, width(1000) replace  
(file fires.png written in PNG format)



Forest Fires At Random Times

# Poisson Regression

. clear all

. set maxvar 10000

. use "/Users/agrogan/Box Sync/DATA WAREHOUSE/General Social Survey/GSS7218\_R1.DTA", clea  
> r

. codebook numprobs // data from 2002  
  
─────────────────────────────────────────────────────────────────────────────────────────  
numprobs how many friends close to discuss problems   
─────────────────────────────────────────────────────────────────────────────────────────  
  
 type: numeric (byte)  
 label: LABJP, but 33 nonmissing values are not labeled  
  
 range: [0,96] units: 1  
 unique values: 34 missing .: 0/64,814  
 unique mv codes: 3 missing .\*: 62,141/64,814  
  
 examples: .i IAP  
 .i IAP  
 .i IAP  
 .i IAP

. recode hispanic (1=0)(2/max = 1), generate(hisp) // dichotomous variable  
(26650 differences between hispanic and hisp)

. generate coninc\_10K = coninc / 10000 // $10K chunks  
(6,520 missing values generated)

. poisson numprobs coninc\_10K i.race hisp i.class sex age  
  
Iteration 0: log likelihood = -13723.05   
Iteration 1: log likelihood = -13723.046   
Iteration 2: log likelihood = -13723.046   
  
Poisson regression Number of obs = 2,399  
 LR chi2(9) = 1125.86  
 Prob > chi2 = 0.0000  
Log likelihood = -13723.046 Pseudo R2 = 0.0394  
  
───────────────┬────────────────────────────────────────────────────────────────  
 numprobs │ Coef. Std. Err. z P>|z| [95% Conf. Interval]  
───────────────┼────────────────────────────────────────────────────────────────  
 coninc\_10K │ .0288957 .0017165 16.83 0.000 .0255315 .0322599  
 │  
 race │  
 black │ -.4409339 .0256346 -17.20 0.000 -.4911767 -.390691  
 other │ -.2608266 .037588 -6.94 0.000 -.3344978 -.1871555  
 │  
 hisp │ -.2247009 .033778 -6.65 0.000 -.2909047 -.1584972  
 │  
 class │  
working class │ -.1267131 .0342909 -3.70 0.000 -.1939219 -.0595042  
 middle class │ -.0173452 .0349358 -0.50 0.620 -.0858182 .0511278  
 upper class │ -.0300967 .049953 -0.60 0.547 -.1280028 .0678095  
 │  
 sex │ .1291068 .0146144 8.83 0.000 .1004631 .1577506  
 age │ -.0013719 .0004456 -3.08 0.002 -.0022453 -.0004986  
 \_cons │ 1.935927 .0457132 42.35 0.000 1.846331 2.025523  
───────────────┴────────────────────────────────────────────────────────────────

# Incidence Rate Ratios

. poisson, irr  
  
Poisson regression Number of obs = 2,399  
 LR chi2(9) = 1125.86  
 Prob > chi2 = 0.0000  
Log likelihood = -13723.046 Pseudo R2 = 0.0394  
  
───────────────┬────────────────────────────────────────────────────────────────  
 numprobs │ IRR Std. Err. z P>|z| [95% Conf. Interval]  
───────────────┼────────────────────────────────────────────────────────────────  
 coninc\_10K │ 1.029317 .0017668 16.83 0.000 1.02586 1.032786  
 │  
 race │  
 black │ .6434353 .0164942 -17.20 0.000 .611906 .6765892  
 other │ .7704145 .0289583 -6.94 0.000 .7156974 .8293148  
 │  
 hisp │ .7987551 .0269804 -6.65 0.000 .7475869 .8534254  
 │  
 class │  
working class │ .8809864 .0302098 -3.70 0.000 .8237222 .9422316  
 middle class │ .9828044 .0343351 -0.50 0.620 .9177611 1.052457  
 upper class │ .9703517 .048472 -0.60 0.547 .8798509 1.070161  
 │  
 sex │ 1.137812 .0166285 8.83 0.000 1.105683 1.170874  
 age │ .998629 .000445 -3.08 0.002 .9977572 .9995015  
 \_cons │ 6.930466 .3168141 42.35 0.000 6.336526 7.580077  
───────────────┴────────────────────────────────────────────────────────────────  
Note: \_cons estimates baseline incidence rate.

# Negative Binomial Distribution

## Over-Dispersion

Due to population heterogeneity (diversity, variation), variance may be mean. This is often empirically the case.

## Regression

. nbreg numprobs coninc\_10K i.race hisp i.class sex age  
  
Fitting Poisson model:  
  
Iteration 0: log likelihood = -13723.05   
Iteration 1: log likelihood = -13723.046   
Iteration 2: log likelihood = -13723.046   
  
Fitting constant-only model:  
  
Iteration 0: log likelihood = -7550.8728   
Iteration 1: log likelihood = -7534.4222   
Iteration 2: log likelihood = -7534.4131   
Iteration 3: log likelihood = -7534.4131   
  
Fitting full model:  
  
Iteration 0: log likelihood = -7461.8838   
Iteration 1: log likelihood = -7458.4673   
Iteration 2: log likelihood = -7458.4576   
Iteration 3: log likelihood = -7458.4576   
  
Negative binomial regression Number of obs = 2,399  
 LR chi2(9) = 151.91  
Dispersion = mean Prob > chi2 = 0.0000  
Log likelihood = -7458.4576 Pseudo R2 = 0.0101  
  
───────────────┬────────────────────────────────────────────────────────────────  
 numprobs │ Coef. Std. Err. z P>|z| [95% Conf. Interval]  
───────────────┼────────────────────────────────────────────────────────────────  
 coninc\_10K │ .0286549 .0049747 5.76 0.000 .0189047 .0384052  
 │  
 race │  
 black │ -.4378139 .0599275 -7.31 0.000 -.5552697 -.320358  
 other │ -.248534 .0944197 -2.63 0.008 -.4335933 -.0634747  
 │  
 hisp │ -.2523914 .0861951 -2.93 0.003 -.4213307 -.083452  
 │  
 class │  
working class │ -.1323204 .0890544 -1.49 0.137 -.3068639 .042223  
 middle class │ -.0193732 .091597 -0.21 0.832 -.1989 .1601535  
 upper class │ -.066233 .1389252 -0.48 0.634 -.3385213 .2060554  
 │  
 sex │ .1082123 .0393212 2.75 0.006 .0311442 .1852805  
 age │ -.0013676 .0011595 -1.18 0.238 -.0036401 .0009049  
 \_cons │ 1.975325 .1187731 16.63 0.000 1.742534 2.208116  
───────────────┼────────────────────────────────────────────────────────────────  
 /lnalpha │ -.246852 .0319159 -.309406 -.184298  
───────────────┼────────────────────────────────────────────────────────────────  
 alpha │ .7812563 .0249345 .7338827 .8316879  
───────────────┴────────────────────────────────────────────────────────────────  
LR test of alpha=0: chibar2(01) = 1.3e+04 Prob >= chibar2 = 0.000

## Predicted Values

## Exposure

In some data sets, we will have a *years exposed* or *time exposed* variable. It is important to control for this variable.

# Zero-Inflated Models